

La détection des patrons de réponses problématiques dans le contexte des tests informatisés¹

Jean-Guy Blais, Université de Montréal

Gilles Raïche, Université du Québec à Montréal

David Magis, Université de Liège

1. Introduction

Le résultat obtenu par une personne à un test n'est pas toujours représentatif de son niveau réel d'habileté et peut constituer une information trompeuse sur ses capacités. Il y a des cas où les candidats manquent d'attention, de motivation ou de préparation, mais il y a également des situations où des personnes performant intentionnellement en deçà de leurs capacités à des tests de classement en vue, par la suite, de fournir un effort moindre qui ne correspond pas à leur vrai potentiel. À titre d'exemple, c'est la situation qu'on observe à l'intérieur du réseau collégial au Québec au regard du test de classement en anglais, langue seconde (Raïche, 2004). Dans d'autres situations, les réponses peuvent être copiées sur celles d'une autre personne et le résultat obtenu est alors une mauvaise estimation du niveau d'habileté, celui-ci se trouvant surestimé. Pour chacun de ces contextes, cela se traduit par la production d'un patron de réponses qui ne permet pas de connaître le niveau d'habileté réel de l'étudiant.

Dans le cas des tests conventionnels, papier-crayon ou écran-clavier, de longueur et de contenu fixes, des indices d'ajustement ont été proposés pour détecter des patrons

¹ La production de ce chapitre a pu être réalisée grâce à l'obtention de subventions auprès du Conseil de recherche en sciences humaines du Canada (CRSH), du Fonds québécois de la recherche sur la société et la culture (FQRSC), d'une bourse de stage doctoral de l'Université de Liège et du Programme d'aide aux chercheurs et chercheuses de l'Université du Québec à Montréal (PAFACC).

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

de réponses dits «problématiques». Ces indices s'avèrent très efficaces lorsque le niveau de difficulté des items qui composent le test varie suffisamment au dessus et au dessous du niveau d'habileté réel du candidat. Toutefois, dans le cas de l'administration de tests adaptatifs par ordinateur, forme de tests sur mesure qui se modifient en fonction des réponses fournies par le candidat (voir le chapitre de Boiteau et Paré dans ce volume), la stratégie du «sur mesure» consiste à faire en sorte que le niveau de difficulté des items se rapproche constamment du niveau d'habileté de l'étudiant. Ainsi, le niveau de difficulté moyen des items soumis à une même personne pourrait varier très peu et les indices de détection des patrons de réponse problématiques s'avèrent peu efficaces dans ces situations. Malgré cela, est-il tout de même possible d'utiliser des indices de détection de réponses problématiques dans le contexte des tests adaptatifs? Avec comme objectif de contribuer à fournir des réponses à ces questions, ce chapitre présente des propositions de modélisations des paramètres associés aux personnes (plutôt qu'aux items) qui pourraient être pertinents pour la détection des patrons de réponses problématiques dans le contexte des tests adaptatifs par ordinateur.

2. Le contexte

Un test adaptatif est un test où la séquence d'administration des items peut se modifier en fonction des réponses obtenues à des items préalablement administrés. Les items administrés et leur nombre peuvent donc être différents d'une personne à une autre. On évite donc l'administration d'items trop faciles à un candidat dont le niveau d'habileté est élevé, ou d'items trop difficiles lorsque le niveau d'habileté du candidat est plutôt faible. Cette stratégie permet alors de diminuer la longueur du test et d'augmenter la

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

précision de la mesure du niveau d'habileté de l'étudiante ou de l'étudiant (Thissen & Mislevy 2000, Boiteau & Paré 2008).

Un des problèmes rencontrés avec les tests adaptatifs est la difficulté de détecter un patron de réponses qui serait problématique. En effet, la détection d'un tel patron de réponses a été bien étudié dans le cas d'un test conventionnel et plusieurs auteurs ont proposé des stratégies et indices à cette fin (Drasgow, Levine & Zickar, 1996; Molenaar & Hoijtink, 2000; Reise & Flannery, 1996; Raîche, 2002; Raîche & Blais, 2003). Ces indices sont particulièrement efficaces lorsque le niveau de difficulté des items constituant le test varie considérablement. On peut ainsi détecter si une personne rate des items qui devraient être faciles pour elle, ou encore, réussit des items qui devraient être trop difficiles. Dans le cas d'un test adaptatif cependant, il est nécessaire de minimiser la variation du niveau de difficulté des items. Au fil de l'administration d'un test adaptatif, le niveau de difficulté des items se rapproche alors constamment du niveau d'habileté de l'étudiant. Or cette stratégie de minimisation va à l'encontre de la stratégie de maximisation de la variation du niveau de difficulté des items qui doit être utilisée pour détecter un patron de réponses qui pourrait être problématique. Comment peut-on alors tenir compte de cet obstacle dans la mise en œuvre d'un test adaptatif?

3. Les indices de détection des patrons de réponses problématiques

Au départ, une des premières études publiées sur ce sujet, celle de Nering (1997), s'intéressait à deux indices, I_z et $ECI4_z$, qui ont été appliqués auparavant avec succès à des données provenant de la passation de tests conventionnels avec le format papier-crayon. Cependant, Nering soulignait tout de même le faible pouvoir de détection de ces indices et la difficulté d'interpréter les valeurs calculées dans le contexte d'un test

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

adaptatif étant donné que la distribution de probabilité de ces indices s'éloigne considérablement d'une loi de probabilité normale, limitant les inférences possibles sous la forme d'intervalles de confiance classiques. Il faut toutefois noter, que ce constat est le même dans le contexte des tests conventionnels, mais qu'il est cependant moins sérieux. McLeod et Lewis (1999) ont également étudié l'efficacité des indices I_z et $EClA_z$, mais, l'originalité de leur travail réside dans l'utilisation d'un indice supplémentaire qui aurait pu s'avérer fort approprié au contexte d'un test adaptatif, il s'agit de l'indice Z_c . Cet indice a la particularité de subdiviser le niveau de difficulté des items administrés en trois catégories : les items faciles, les items de difficulté moyenne et les items difficiles. Toutefois, selon les résultats présentés par les auteurs, cet indice s'est avéré fort peu efficace. Pour leur part, van Krimpen-Stoop et Meijer (1999b), ont utilisé un indice de détection I_z^* tenant compte du biais dans l'estimation du niveau d'habileté. D'ailleurs, on peut avancer que cet indice est actuellement l'indice le plus recommandé dans le contexte des tests conventionnels. À partir de cet indice, van Krimpen-Stoop et Meijer n'ont cependant pas pu arriver à des résultats plus intéressants que ceux obtenus auparavant par Nering et par McLeod et Lewis dans le cas des tests adaptatifs. Ils ont fait face aux mêmes problèmes d'interprétation et d'efficacité de détection.

Par la suite, van Krimpen-Stoop et Meijer (1999a, 2000), ainsi que Meijer (2002), ont réalisé quelques études similaires avec des indices dédiés au contexte des tests adaptatifs. Ces indices sont issus des techniques de contrôle de la qualité utilisées en industrie (Page, 1954; Stoumbos et Reynolds, 2001). Toutefois, le problème de l'interprétation des valeurs de ces indices demeure même si l'efficacité de la détection était meilleure que dans les tentatives précédentes. La dernière recherche recensée, celle

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

réalisée par Meijer (2004), propose l'utilisation d'un indice non paramétrique basé sur le score total à des sous-tests, l'indice L_{xx} . Encore une fois, les résultats obtenus amènent à conclure que cet indice est très peu efficace pour détecter des patrons de réponse problématiques dans le contexte de l'administration de tests adaptatifs. Intuitivement, il nous apparaît que ce peu d'efficacité soit dû à un problème d'interprétation de la distribution de probabilité de l'indice.

Raïche (2002), ainsi que Raïche et Blais (2003), ont proposé une solution au problème d'interprétation de la distribution de probabilité des indices de détection dans le contexte des tests conventionnels : ils ont utilisé la distribution empirique (i.e. la distribution observée) des indices de détection pour en interpréter les valeurs, plutôt qu'une distribution théorique déterminée à l'avance. Cette stratégie a permis d'augmenter considérablement la puissance de détection des indices étudiés dans ces travaux, soit I_z , *Zeta*, *Infit* et *Outfit*. À prime abord, l'application de cette stratégie dans le contexte des tests adaptatifs semble prometteuse. Elle présente toutefois le désavantage d'exiger un temps de calcul considérable. D'autres solutions, où on se préoccupe du contexte d'application des indices, sont aussi à explorer. Par exemple, tenir compte du fait que les comportements qui produisent les patrons de réponses problématiques dans un test adaptatif peuvent être différents de ceux à l'œuvre dans un test conventionnel; utiliser un estimateur du niveau d'habileté moins biaisé; augmenter l'étendue du niveau de difficulté des items administrés ou encore utiliser des indices de détection qui conviendraient à des comportements spécifiques, comme l'inattention ou les réponses données au hasard.

La suite de ce chapitre propose ainsi une approche quelque peu différente pour caractériser le degré d'adéquation du patron de réponse d'une personne, qu'il soit

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

adaptatif ou non. Les modélisations étudiées pourraient permettre de surmonter certaines limites des approches de détection des patrons de réponses problématiques utilisées jusqu'à maintenant. Ainsi, il sera éventuellement possible d'obtenir une mesure du degré d'adéquation du patron de réponses qui n'est pas tributaire du biais quant au niveau d'habileté estimé. En fait, la plupart des indices d'adéquation sont calculés à partir d'un estimateur du niveau d'habileté qui est biaisé et le biais est encore plus marqué lorsqu'une méthode d'estimation bayésienne est utilisée. L'estimation simultanée du degré d'adéquation et du niveau d'habileté à l'intérieur d'une même modélisation pourrait éventuellement diminuer l'impact d'un tel biais. De plus, actuellement les différents indices d'adéquation d'un patron de réponses ne permettent pas de leur associer une erreur type et ainsi d'estimer leur précision. À partir des modélisations proposées dans ce chapitre, on pourra éventuellement produire une mesure ou des mesures d'adéquation pour lesquelles il est possible d'obtenir directement un estimateur de la précision.

4. L'intégration de paramètres associés aux personnes

De façon générale, les modèles de la théorie des réponses à l'item intègrent des paramètres qui sont associés aux personnes et aux items. Toutefois, lorsque des paramètres sont ajoutés aux modèles plus simples pour décrire la réalité, ils ont le plus souvent comme fonction de modéliser plus adéquatement les caractéristiques des items et non celles des personnes.² Un bon exemple à cet effet dans la situation où les items sont à réponse choisie est le modèle de Rasch qui devient le modèle logistique avec deux paramètres lorsqu'on ajoute un paramètre pour tenir compte du pouvoir de discrimination

² Les modèles multidimensionnels peuvent aussi être considérés comme des modèles auxquels on ajoute des paramètres, mais ils ne seront pas abordés dans ce texte.

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

de l'item, et qui devient le modèle avec trois paramètres lorsqu'on ajoute un paramètre pour tenir compte de la possibilité qu'une personne devine la bonne réponse (paramètre de pseudo-chance). Dans ce dernier cas, même si c'est la personne qui devine, la modélisation considère qu'il s'agit d'une propriété inhérente à l'item, celui-ci favorisant par ses caractéristiques le choix au hasard de la réponse par certaines personnes.

Pour illustrer le développement de modèles où on s'intéresse aux caractéristiques «métriques» des personnes plutôt qu'à celles des items, un modèle unidimensionnel simple, le modèle de Rasch pour la situation où la réponse est dichotomique, a été retenu et il sera complexifié par l'ajout de paramètres «personnels» tentant de rendre compte de différents comportements des répondants. Ainsi, la modélisation d'une bonne réponse X_{ij} de la part d'un candidat j à un item i , conditionnelle au paramètre représentant la difficulté de l'item, b_i , et à un premier paramètre personnel du candidat, soit son niveau d'habileté θ_j , prend la forme de l'équation ci-dessous.

$$P_{ij} = P(X_{ij} = 1 | \theta_j, b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}}.$$

Par analogie avec ce qui se passe lorsque les paramètres ajoutés visent les items, de nouveaux paramètres pour les personnes, celui de discrimination α_j , celui de pseudo chance χ_j et celui d'inattention δ_j , sont ajoutés simultanément au niveau d'habileté θ_j et le modèle obtenu correspondrait à l'équation suivante.

$$P(x_{ij} = 1 | \theta_j, \alpha_j, \chi_j, \delta_j, b_i) = \chi_j + \frac{\delta_j - \chi_j}{1 + e^{-\alpha_j(\theta_j - b_i)}}.$$

En se limitant à un exemple simple comportant peu d'items, on supposera deux patrons de réponses hypothétiques obtenus par deux candidats ($j=1,2$) à 9 items ($i=1,\dots,9$), que nous utiliserons pour comparer les différentes modélisations (une valeur de 1 correspondant à une bonne réponse à l'item i et une valeur de 0 à une mauvaise réponse):

$$x_{i1} = \{1,1,1,1,1,0,0,0\}$$

et

$$x_{i2} = \{1,0,1,1,1,1,0,0,1\},$$

Supposons un ensemble de valeurs associées au paramètre b_j pour ces 9 items : $\{-2, 0, 0, 0, 1, 2, 3, 3, 3\}$. Les valeurs ayant été placées en ordre ascendant, il est aisé de leur faire correspondre les réponses des candidats à chacun des items. Ainsi, on note que le candidat 1 présente un patron de réponses «adéquat» au sens de Guttman. Il a donné de bonnes réponses pour les six premiers items et de mauvaises réponses pour les trois derniers items; la hiérarchie de la difficulté des items est donc respectée. Le candidat 2, pour sa part, présente un patron de réponses où il obtient une mauvaise réponse aux items 2, 7 et 8, et une réponse correcte à l'item 9. Son patron de réponse est plutôt inusité, même s'il s'agit d'un patron de réponses réaliste.

En appliquant la méthode d'estimation par maximum de vraisemblance a posteriori avec le modèle de Rasch pour les réponses dichotomiques, on obtient un niveau d'habileté θ_1 égal à 1,3 pour le premier candidat. Le second candidat, quant à lui, a obtenu le même nombre de bonnes réponses et on devrait obtenir avec ce modèle une estimation du niveau d'habileté égale à la valeur obtenue pour le premier candidat. C'est

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

le cas, et on obtient la même valeur 1,3 pour le paramètre θ_2 . La figure 1 représente la relation entre le niveau d'habileté estimé et la probabilité d'obtention du patron de réponses.

Insérer la figure 1

Nous allons maintenant comparer ces résultats à ceux qu'on obtient avec des modélisations où d'autres paramètres pour les personnes sont ajoutés.

De la même façon qu'on produit ce qu'on appelle la courbe caractéristique d'un item (voir le chapitre 2 dans ce livre), il est possible de produire la courbe caractéristique d'un candidat. Plusieurs appellations ont été suggérées pour cette représentation graphique : courbe de trace (*trace line*, Weiss 1973), courbe caractéristique de personne (*person characteristic curve*, Lumsden 1977, 1978), courbe de réponse de personne (*person response curve*, Trabin et Weiss 1983). Nous avons retenu la deuxième appellation et nous utiliserons donc dans ce qui suit l'expression «courbe caractéristique de la personne» (CCP).

La figure 2 offre une représentation de trois courbes caractéristiques d'une personne pour trois candidats hypothétiques (exemple adapté de Trabin et Weiss, 1983, p. 91). Supposons qu'on administre un assez grand nombre d'items à ces trois candidats, plus de 200 par exemple, et qu'on regroupe ces items en sept classes de niveaux de difficulté. Il est alors possible de calculer avec quelle fréquence relative chacun des candidats a donné une bonne réponse aux items de chacune des catégories. Pour le premier candidat, on observe une courbe caractérisant une personne qui obtient à coups

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

sûrs une bonne réponse lorsque le niveau de difficulté des items est faible et une mauvaise réponse de façon assurée lorsque les items sont d'un niveau de difficulté élevé. Entre les items de niveau de difficulté faible et ceux de niveau de difficulté élevée, la courbe est graduellement descendante et affiche une forme en S typique de ce qui est attendu avec le modèle de Rasch pour des données dichotomiques.

Les deuxième et troisième candidats affichent toutefois des courbes caractéristiques atypiques. Ainsi, ces candidats ne réussissent pas à coup sûr les items les plus faciles, comme s'ils pouvaient être caractérisés par un paramètre personnel d'inattention plus faible que la valeur attendue. La valeur du paramètre personnel δ_j de ces candidats serait alors inférieure à la valeur attendue de 1. Dans le cas du deuxième candidat, on observe que lorsque des items d'un niveau de difficulté élevé lui sont administrés il réussit ceux-ci plus fréquemment que prévu par la modélisation. Un paramètre personnel de pseudo chance χ_2 avec une valeur supérieure à 0 pourrait donc être en jeu. Deuxièmement, on observe une remontée inattendue de la fréquence relative de réponses correctes au centre de la courbe. On pourrait alors émettre deux hypothèses qui impliquent toutes deux l'ajout de paramètres associés à la personne supplémentaires. La première hypothèse serait qu'au moins un autre niveau d'habileté serait impliqué et qu'on aurait plutôt affaire à un vecteur de niveaux d'habileté $\theta = \{\theta_1 \dots \theta_n\}$, i.e. une situation multidimensionnelle. La seconde hypothèse serait qu'un paramètre personnel de discrimination α_j serait en jeu. Cette dernière hypothèse peut être mise en relation facilement avec la première, car une telle modification de la discrimination dans le cas d'un item est d'ailleurs expliquée par la multidimensionalité des niveaux d'habileté (McDonald, 1981).

Insérer la figure 2

De plus, le paramètre de discrimination, qu'il soit associé à un item ou à un candidat, correspond à l'inverse de la racine carrée d'une variance. Ce qui veut dire que plus la valeur du paramètre de discrimination est faible, plus la variabilité du niveau de difficulté de l'item ou, comme ici, du niveau d'habileté du sujet, est grande. C'est ce qu'ont souligné Levine et Drasgow (1983) en proposant une nouvelle modélisation de la réponse à l'item qui intègre, en plus du niveau d'habileté, un paramètre qui modélise la variabilité du niveau d'habileté du sujet à l'intérieur d'un même test. La modélisation repose sur une simple modification de la formulation du modèle de Rasch pour des données dichotomiques :

$$P(x_{ij} = 1 | \theta_j, \alpha_j, b_i) = \frac{1}{1 + e^{-\alpha_j(\theta_j - b_i)}},$$

Un paramètre de discrimination associé à la personne, α_j , a ainsi été ajouté au modèle de Rasch. Peu de travaux ont été réalisés au regard de cette modélisation jusqu'à récemment. Les recherches encore exploratoires de Ferrando (2004), de Magis (2007) et de Magis et Raïche (2007) ont permis de prendre en considération de nouveau cette

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

modélisation. En poursuivant avec l'exemple ci-dessus, la méthode du maximum de vraisemblance a posteriori permet d'obtenir les valeurs suivantes : $\theta_1 = 1,58$; $\alpha_1 = 1,78$; $\theta_2 = 1,30$; $\alpha_2 = 0,98$. Ainsi, pour le premier candidat, celui dont le patron de réponses est conforme à une échelle de Guttman, le niveau d'habileté est supérieur à celui obtenu à partir de la modélisation de Rasch qui était de 1,30 et la discrimination personnelle devient supérieure à celle prévue de 1. Dans le cas du deuxième candidat, le niveau d'habileté est le même, tandis que le paramètre de discrimination personnelle est plus faible que la valeur attendue de 1. Pour cette personne, comme on pouvait s'y attendre, la discrimination est plus faible. Les figures 3a et 3b illustrent la probabilité associée à différentes valeurs des estimateurs des paramètres personnels des deux premiers candidats. On y remarque bien la correspondance avec les valeurs que nous avons obtenues à partir de la méthode d'estimation par vraisemblance maximale a posteriori.

Insérer la figure 3

Les explications accompagnant la dérivation de cette équation par Levine et Drasgow (1983) sont instructives en ce sens qu'elles montrent que celle-ci ne peut pas malheureusement être utilisée pour corriger la valeur du niveau d'habileté malgré la valeur obtenue aux autres paramètres d'items, car on obtient plutôt un estimateur moyen du niveau d'habileté à ce test et non pas un estimateur fixe du niveau d'habileté du sujet. Ce qu'on peut toutefois obtenir c'est une information supplémentaire sur la variabilité de ce niveau d'habileté qu'on devrait alors décrire comme une variable aléatoire présentant

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

une variance ($N(\theta_j, \alpha_j^2)$) et non pas simplement comme une valeur fixe du niveau d'habileté. Plus de travaux sont maintenant à réaliser au regard de l'interprétation des paramètres personnels dans cette modélisation.

Les deux modélisations suivantes prennent en compte un paramètre personnel de pseudo-chance χ_j et un paramètre personnel d'inattention δ_j (plutôt qu'un paramètre personnel additionnel de discrimination) :

$$P(x_{ij} = 1 | \theta_j, \chi_j, b_i) = \chi_j + \frac{1 - \chi_j}{1 + e^{-(\theta_j - b_i)}}$$

$$P(x_{ij} = 1 | \theta_j, \delta_j, b_i) = \frac{\delta_j}{1 + e^{-(\theta_j - b_i)}}$$

Dans le cas de la modélisation intégrant un paramètre personnel de pseudo chance, les estimations obtenues pour les deux candidats hypothétiques correspondent à : $\theta_1 = 1,02$; $\chi_1 = 0,18$; $\theta_2 = 0,08$; $\chi_2 = 0,49$. Dans les deux cas le niveau d'habileté estimé est plus faible, et d'autant plus faible que le paramètre personnel de pseudo chance est élevé. Les figures 4a et 4b illustrent la vraisemblance des différentes valeurs des estimateurs des paramètres personnels des deux candidats. Il faut noter que les fonctions utilisées imposent logiquement une limite inférieure de 0 et une limite supérieure de 1 au paramètre personnel de pseudo chance. Il en sera de même pour le paramètre d'inattention plus loin.

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

Dans ce cas, comme dans tous ceux impliquant les paramètres personnels de pseudo chance ou d'inattention, la fonction de maximisation que nous avons utilisée a permis de tenir compte uniquement de la distribution de probabilité a priori du niveau d'habileté. Les distributions de probabilité a priori (distributions beta) des paramètres personnels de pseudo chance ou d'inattention que nous avons expérimentées se sont avérées peu stables et donnaient des résultats imprévisibles.

Insérer la figure 4

Lors de l'ajout d'un paramètre personnel relié à l'inattention, δ_j , tel qu'attendu avec le modèle de Rasch la valeur du paramètre d'inattention est égale à 1 pour les deux patrons de réponses considérés dans l'exemple. L'estimateur du niveau d'habileté demeure lui aussi inchangé avec une valeur de 1,30. Les figures 5a et 5b illustrent la variation de la vraisemblance des estimations des paramètres personnels d'inattention.

Insérer la figure 5

Comme nous l'avons mentionné, il est aussi possible de penser à des modélisations plus complexes intégrant trois ou quatre paramètres personnels. La représentation graphique des fonctions de vraisemblance associées ne sera pas considérée à partir de maintenant, car elle est difficile à réaliser, quatre dimensions étant maintenant

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

impliquées, cinq dans le cas de du modèle qui intègre quatre paramètres associés aux personnes.

Avec un modèle intégrant des paramètres personnels de discrimination et d'inattention (voir ci-dessous), les valeurs obtenues sont les suivantes : $\theta_1 = 1,58$; $\alpha_1 = 1,78$; $\delta_1 = 1,00$; $\theta_2 = 1,30$; $\alpha_2 = 0,98$; $\delta_2 = 1,00$. Puisque le paramètre personnel d'inattention n'est pas affecté, nous retrouvons pour ces deux patrons de réponse spécifiques le même résultat que celui obtenu à partir de la modélisation intégrant uniquement le paramètre personnel supplémentaire de discrimination.

$$P(x_{ij} = 1 | \theta_j, \alpha_j, \delta_j, b_i) = \frac{\delta_j}{1 + e^{-\alpha_j(\theta_j - b_i)}}$$

En ce qui a trait à la modélisation intégrant simultanément les paramètres personnels de discrimination et de pseudo chance (équation ci-dessous), on retrouve encore les mêmes paramètres personnels pour le premier candidat, puisque le paramètre personnel de pseudo chance est nul. Toutefois, pour le deuxième candidat, on remarque une modification importante des paramètres personnels de discrimination et de pseudo chance. Les valeurs des paramètres estimés sont les suivantes : $\theta_1 = 1,58$; $\alpha_1 = 1,78$; $\chi_1 = 0,00$; $\theta_2 = 0,00$; $\alpha_2 = 1,39$; $\chi_2 = 0,52$.

$$P(x_{ij} = 1 | \theta_j, \alpha_j, \chi_j, b_i) = \chi_j + \frac{1 - \chi_j}{1 + e^{-\alpha_j(\theta_j - b_i)}}$$

Enfin, le dernier modèle intègre tous les paramètres personnels que nous avons considérés jusqu'à maintenant. Dans ce cas précis, puisque la valeur du paramètre

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

personnel d'inattention demeure inchangée et est égale à 1, nous retrouvons exactement les mêmes valeurs pour les paramètres personnels que celles avaient été obtenues avec la modélisation précédente. Il est évidemment possible que d'autres patrons de réponses produisent des résultats différents.

$$P(x_{ij} = 1 | \theta_j, \alpha_j, \chi_j, \delta_j, b_i) = \chi_j + \frac{\delta_j - \chi_j}{1 + e^{-(\theta_j - b_i)}}$$

5. Conclusion et implications pour les tests adaptatifs

Nous avons tenté de proposer une approche assez différente pour caractériser le degré d'adéquation du patron de réponse d'un sujet à un test, qu'il soit adaptatif ou non. À cette fin, l'ajout de paramètres personnels à la modélisation logistique à un paramètre de Rasch a été considéré. Ainsi, trois paramètres personnels ont été pris en compte, soit les paramètres personnels de discrimination, de pseudo chance et d'inattention.

La modélisation où uniquement le paramètre personnel de discrimination est ajouté semble la plus intéressante pour la détection de patrons de réponses problématiques à l'intérieur de tests adaptatifs. L'estimation de ce paramètre personnel additionnel semble assez facile à obtenir avec une méthode bayésienne comme celle de maximisation a posteriori. Toutefois, comme le souligne Ferrando (2004), pour obtenir un estimateur précis le double du nombre d'items nécessaires pour obtenir un estimateur précis du niveau d'habileté est requis. Cela peut entraîner le besoin de plus de 80 items. Toutefois dans un test adaptatif, on pourrait utiliser le paramètre personnel de discrimination pour optimiser le choix des prochains items à administrer et

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

éventuellement obtenir un estimateur moins biaisé du niveau d'habileté. Malgré cette opportunité, il est possible que l'estimateur du niveau d'habileté n'ait plus la même signification que celui obtenu à partir de la modélisation logistique usuelle à un paramètre de Rasch. Cette remarque est valide pour toutes les modélisations alternatives que nous avons proposées ici. Il ne semble donc pas possible, malheureusement, de corriger l'estimateur du niveau d'habileté du sujet malgré le fait que son patron de réponses soit problématique.

Toutes les modélisations alternatives où des paramètres personnels additionnels de pseudo chance ou d'inattention sont ajoutés, quoique potentiellement intéressantes d'un point de vue théorique, ne semblent pas très utiles pour le moment. Premièrement, les estimateurs exigent que des adaptations soient effectuées au regard des méthodes d'estimation. Actuellement, les estimateurs obtenus ne semblent pas vraiment stables. De plus, comme pour leur estimation avec les items, nécessitent plusieurs items très faciles ou très difficiles pour que les valeurs obtenues aient un sens. Il est probable que pour qu'on puisse obtenir des estimateurs d'une précision satisfaisante, il soit nécessaire que le patron de réponses du sujet soit constitué encore de plus d'items que dans le cas du paramètre personnel de discrimination.

Une des applications les plus intéressantes de toutes ces modélisations sera sans contredit leur utilisation à l'intérieur de simulations de tests classiques fixes et invariables et à l'intérieur de tests adaptatifs. Il sera ainsi très facile de contrôler le degré d'inadéquation de patrons de réponses simulées en faisant varier les paramètres

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

personnels. Dans le cas des tests adaptatifs, cela s'avère encore plus utiles, car il a toujours été difficile de simuler adéquatement des patrons de réponses inadéquats.

Une dernière application de ces travaux pourra consister à utiliser ces modélisations, surtout celle où un paramètre personnel additionnel de discrimination est ajouté, pour effectuer la calibration simultanée des paramètres d'items et des paramètres personnels. Il ne serait éventuellement plus nécessaire d'identifier dans un premier temps les sujets dont le patron de réponses est problématique et de les retirer ensuite du processus de calibration. On pourrait ainsi tenir compte directement de leur degré d'inadéquation.

6. Références

- Drasgow, F., Levine, M.V. et Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Psychological Measurement*, 9(1), 47-64.
- Ferrando, P. J. (2004). Person reliability in personality measurement : an item response theory analysis. *Applied Psychological Measurement*, 28(2), 126-140.
- Levine, M. V. et Drasgow, F. (1983). Appropriateness measurement: validating studies and variable ability models. Dans D. J. Weiss (Dir.) : *New horizons in testing. Latent trait test theory and computerized adaptive testing*. New York : Academic Press.
- Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, 1, 477-482.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31, 19-26.

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

Magis, D. (2007). *Influence, information and item response theory in discrete data analysis*. Thèse de doctorat inédite. Liège, Belgique : Université de Liège.

Magis, D. et Raïche, G. (2007). *Two new iterative Bayesian estimators of subject ability and their properties*. Rapport no 07.001. Département de mathématiques, Université de Liège, Liège, Belgique.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

McLeod, L.D. et Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, 23(2), 147-160.

Meijer, R.B. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, 39(3), 219-233.

Meijer, R.B. (2004). Using patterns of summed scores in paper-and-pencil tests and computer-adaptive tests to detect misfitting item score patterns. *Journal of Educational Measurement*, 41(2), 119-136.

Molenaar, I.W. et Hoijsink, H. (2000). The many null distributions of person fit indices. *Psychometrika*, 55(1), 75-106.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, 21(2), 115-127.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100-115.

Raïche, G. (2002). *Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial*. Hull : Collège de l'Outaouais.

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

Raïche, G. (2004). Le testing adaptatif. Dans R. Bertrand et J.-G. Blais, *Modèles de mesure : l'apport de la théorie des réponses aux items*. Sainte-Foy : Presses de l'Université du Québec.

Raïche, G. et Blais, J.-G. (2003). The distribution of person-fit indices conditional on the estimated proficiency level and the detection of underachievement at a placement test. Communication réalisée à la rencontre annuelle de la Psychometric Society à Cagliari, Sardaigne.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reese, S.P. et Flannery, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Psychological Measurement*, 9(1), 9-26.

Stoumbos, Z.G. et Reynolds, M.R. (2001). The state of statistical process control as we proceed into the 21th Century. *Journal of the American Statistical Association*, 95(451), 992-994.

Thissen, D. et Mislevy, R.J. (2000). Testing algorithms. Dans H. Wainer (Éd.): *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Trabin, T. E. et Weiss, D. J. (1983). The person response curve: fit of individual to item response theory models. Dans D. J. Weiss (Dir.) : *New horizons in testing. Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Texte préliminaire qui pourrait être intégré à Blais, J.-G. (2008). *L'utilisation des technologies de l'information pour l'évaluation*. Ste-Foy, Québec : Presses de l'Université Laval.

van Krimpen-Stoop, E.M.L.A. et Meijer, R.R. (1999a). *CUSUM-based person-fit statistics for adaptive testing* (Rapport de recherche 99-05). Enschede, NETHERLANDS: University of Twente.

van Krimpen-Stoop, E. M. L. A. et Meijer, R. R. (1999b). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23(4), 327-345.

van Krimpen-Stoop, E.M.L.A. et Meijer, R.R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. Dans W.J. van der Linden et C.A.W. Glas (Éds): *Computerized adaptive testing: theory and practice*. Dordrecht: Kluwer.

Warm, T.A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, 54, 427-450.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Rapport de recherche no 73-3). Minneapolis, MN : University of Minnesota.